Applying Clustering and Association Rule Learning for Finding Patterns in Herbal Formulae

Verayuth Lertnattee[1] and Sinthop Chomya[2]
[1]Department of Health-related Informatics, [2]Department of Pharmacognosy
Faculty of Pharmacy, Silpakorn University, Maung, Nakorn Pathom, 73000 Thailand
E-mail: verayuths@hotmail.com, verayuth@su.ac.th, sinthop@su.ac.th

***Abstract*— Traditional herbal formulae can be usually characterized by the use of several herbs. Various patterns of combinations from these herbs can be applied on a disease. In this paper, we apply two techniques of data mining, i.e., clustering and association rule learning, for finding patterns of herbal formulae with main category of muscle pain and fatigue and several subcategories. With clustering technique, it facilitates herbal experts to find the set of clusters with appropriated subcategories. The association rule learning is applied on the all formulae and formulae on each cluster of the selected set to find the important patterns of combinations. The results show that data mining techniques are useful for finding patterns in herbal formulae.**

***Index Terms*—herbal formulae, data mining, clustering, association rule learning**

I. Introduction

Origins of many traditional treatments in Thailand can be traced to India. The derivation has been diversified throughout many cultures since then [1]. Herbs are natural products that have been used safely for thousands of years to promote healing in patients. They should be taken with caution, and careful consideration of the dosage recommended. Traditional herbal formulae can be usually characterized by the use of several herbs. Various patterns of combinations from these herbs, can be applied on a disease. According to Thai traditional medicine, herbal formulae can be divided into several categories. Some formulae can be classified as more than one category. The categories are usually based on indications of herbs in formulae. A combination of several herbs causes a formula has several categories. These categories may be arranged in flat and/or hierarchy. When the categories are arranged in flat, several main indications of the herbal formula can be applied to patients. In a complex situation, a formula is classified with one main category (or more) and a set of subcategories under the main category. With observation from human, it is hard to discover the combinational patterns of herbs in formulae. Nowadays, several data mining techniques, i.e., classification, clustering, association rules and, etc., have been developed and applied on several types of data. However, only few research works have applied on herbal information. In this paper, we apply two techniques of data mining, i.e., clustering and association rule learning, for finding patterns of herbal formulae with a main category of muscle pain and fatigue and several subcategories. With clustering technique, it facilitates herbal experts to find the set of clusters with appropriated subcategories. The association rule learning is applied on the all formulae of a main category and formulae on each cluster of the selected set to find the important patterns of combinations.

In the rest of this paper, section II presents herbal formulae for muscle pain and fatigue. The concepts of clustering and association rule learning are given in section III. The experimental settings are described in section IV. In section V, a number of experimental results are given. A conclusion is made in section VI.

II. Herbal Formulae for Muscle Pain and Fatigue

At present, a set of Thai traditional medicinal products has been claimed its indication for relieving the aches of body muscles (Kra-sai in term of Thai traditional medicine). However, components and indications of the formulations are different. With the primary indications of muscle pain and fatigue, several secondary indications can be applied such as laxative, stomachic and carminative, body strength, relief of chronic constipation, tendon pain, blood and win disease and muscle paralyses [2]. It is hard to select the best herbal formula for a patient. The number of registered herbal pharmaceutical products for the muscle pain and fatigue is more than 203 formulae. The total number of unique herbs from these formulae is 447 [3]. The relationship between the components and indications is found on almost all formulae. It is quite reliable according to principles of traditional Thai medicine. With the main indication of muscle pain and fatigue, the formulae can be divided into 5 groups of secondary indications according to components in formulae: 1) element balancer, carminative and appetite 2) blood and air distribution 3) laxative and cathartic 4) muscle tonic and 5) body tonic. However, a formula may be one or more secondary indications.

III. Clustering and Association Rule Learning

Several data mining techniques can be applied on herbal formulae. However, two techniques are taken into account in this work i.e., clustering and association rule learning.

A. Clustering on Herbal Formulae

In opposite to classification which is a supervised learning, clustering is an unsupervised learning which is one of the most useful techniques for Clustering technique enables to produce the smaller and more uniform clusters from a large data set. A large number of clustering algorithms was mentioned in the literature. The major clustering methods

DOI: 01.IJNS.03.02. 96

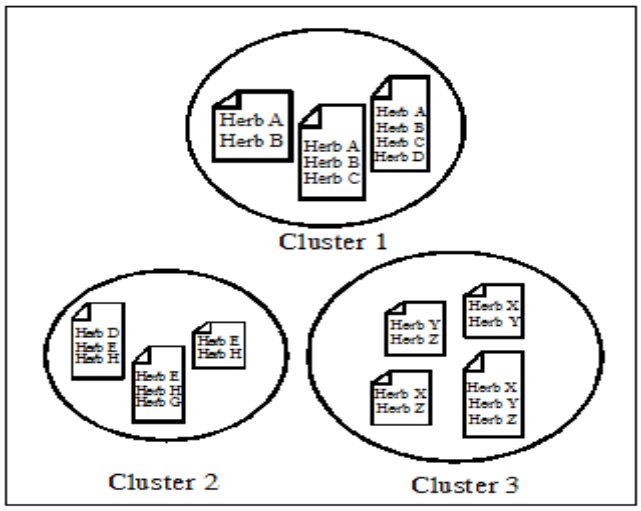

Figure 1. Three clusters of herbal formulae are generated

can be classified into several categories including hierarchical methods [4], density-based methods and model-based methods [5]. The choice of clustering algorithm depends on the particular purpose and application as well as characteristics of data. In this research, a set of herbal formulae for muscle pain and fatigue is collected. We can apply clustering technique to divide the collection of herbal formulae with a specific number of clusters. A set of similar components in formulae should be grouped in the same cluster. In Fig. 1, three clusters are generated based on content of formulae.

B. Association Rule Learning

Association rule learning is a popular and well researched method for finding patterns of interesting associations and correlations between itemsets in large databases. Association rules for discovering regularities between products in large scale transaction data recorded by point-of-sale (POS) systems in supermarkets was introduced by Agrawal et al., [6]. Several research works develop mining techniques on association rule learning such as [7]. In this paper, a set of herbs is considered as a set of items. Each herb has a Boolean value of 0 or 1, here, representing the absence (0) or presence (1) of that herb. Each formula can be represented by a Boolean vector of values assigned to these herbs. The patterns that reflect herbs that are frequently associated or applied together in a formula can be analyzed. These patterns are in the form of association rules. For example, a formula in the main category of muscle pain and fatigue with a subcategory of the laxative, *Herb A* also tends to associate with *Herb B*. This can be represented in association rule below:

$$Herb\ A \rightarrow Herb\ B_{[support=30\%,\ confidence=60\%]}$$

Rule support and confidence are two measures of rule interestingness. They respectively reflect the usefulness and certainty of discovered rules. A support of 30% represents that 30% of all the transactions in all formulae which is used for analyzing, indicate that both *Herb A* and *Herb B* are found together. A confidence of 60% means that 60% of the total number of formulae which is composed of Herb A, is also found Herb B. Typically, association rule learning are considered interesting if they satisfy both a minimum support threshold and a minimum confidence threshold. Such thresholds can be set by users or domain experts. Such information can be used as the basis for decisions about several activities. In this research, a set of rules generated on each clusters are provided useful knowledge about a subcategory while a set of rules from all formulae is informed us the patterns of the main category.

IV. Experimental Setting

To evaluate the concept of applying clustering and association rule learning for finding patterns in herbal formulae, a set of registered pharmaceutical products for muscle pain and fatigue is used. The total number of formulae is 152. The simplest three formulae composed of 3 herbs for each formula. The most complex formula composed of 54 herbs. For a formula, herbs and other materials are then combined. From our preliminary test, when amounts of each herb in a formula are transformed to 1, herbal experts suggested that the result from the Boolean value is better than that of the real value. The 152 formulae are divided into clusters using EM algorithm. The numbers of clusters are 3, 4, 5 and 6. These 4 sets of clusters are evaluated by two herbal experts. To select the best set, the criteria for making decision are as follow.

- Homogeneity of each cluster
- Each cluster should represent a secondary indication of the cluster

The selected set is further explored by the association rule learning.

V. Experimental Results

A. EM Clustering

In this first experiment, EM algorithm is applied to create sets of clusters. The numbers of clusters are set to 3, 4, 5 and 6. The maximum number of iteration is 100. Table I showed the result in forms of the numbers of herbal formulae for each cluster.

TABLE I. The Numbers of Herbal Formulae for Each Cluster

Numbers of Clusters	Clusters					
	0	*1*	*2*	*3*	*4*	*5*
3 Clusters	29	72	51	-	-	-
4 Clusters	18	57	41	36	-	-
5 Clusters	49	51	25	6	21	-
6 Cluster	44	38	12	34	19	5

From the result, experts agree that the set of 5 clusters, is the best. It is more investigated in the next experiment.

B. Association Rule Learning on Herbal Formulae

In this experiment, association rule learning is applied to the all data set and all clusters. The Apriori algorithm is used to generate rules. The support and confidence are set to 30% and 60%, respectively. Table II showed the number of rules, the minimum number of herbs, the maximum number of herbs

DOI: 01.IJNS.03.02.96

and the average number of herbs for a formula on each cluster.

A set of examples of rules generates from all formulae and on each cluster are list below in the format of:
Herb A Herb B→ Herb C (% support, % confidence)

Here, the % support is the percentage of transactions that contain all herbs appearing in the rule, i.e., *Herb A* and *Herb B* and *Herb C*. The % confidence is the confidence of the rule, which is computed as the quotient of the percentage of transactions that contain all herbs appearing in the rule body (antecedent) and the rule head (consequent, i.e., *Herb C*). The herb name is presented by scientific name in italic. The product of herb is presented in regular format.

All formulae
Derris scandens → (77.6%, 77.6%)

Formulae in cluster 0
Derris scandens → (81.6%, 81.6%)
Aloe resin → (67.3%, 67.3%)
Derris scandens → Aloe resin (57.1%, 70.0%)

Formulae in cluster 1
Derris scandens → (70.6%, 70.6%)
Dracaena loureiri → *Cryptolepis buchanani* (51.0%, 72.2%)
Dracaena loureiri → Camphor (19.6%, 62.5%)

Formulae in cluster 2
Derris scandens → (88.0%, 88.0%)
Maerua siamensis → (60.0%, 60.0%)
Senna siamea → (68.0%, 68.0%)
Aloe resin → (72.0%, 72.0%)
Rheum palmatum → (72.0%, 72.0%)

Formulae in cluster 3
Derris scandens → (83.3%, 83.3%)
Cryptolepis buchanani → (66.7%, 66.7%)
Root of *Plumbago indica* → (66.7%, 66.7%)
Angelica sinensis → (66.7%, 66.7%)

TABLE II. THE NUMBERS OF RULES GENERATED AND NUMBERS OF HERBS WITH SUPPORT=30% AND CONFIDENCE=60%

Cluster No.	No. of Rules	Min No. Herb	Max No. Herb	Average No. Herb
All	5	3	54	19
0	18	6	54	19
1	14	3	41	17
2	123	14	53	22
3	16,499	22	31	25
4	3	8	32	17

Dried *Zingiber officinale* → (83.3%, 83.3%)
Angelica dahurica → (83.3%, 83.3%)
Cyperus rotundus → (83.3%, 83.3%)
Piper chaba → (100.0%, 100.0%)

Formulae in cluster 4
Derris scandens → (71.4%, 71.4%)
Derris scandens → *Cryptolepis buchanani* (52.4%, 73.3%)
Cryptolepis buchanani → *Derris scandens* (52.4%, 91.7%)

From the result in Table II and a set of rules generated from herbal formulae, some observations can be summarized as follow. With the support of 30% and confidence of 60%, the numbers of rules are different, especially in cluster 3 which the average number of herbs for a formula is greater than the average numbers of the other clusters. The reason is that the minimum number of herbs of formulae is rather high (22) compare to the other clusters. Huge combinations can be generated from components of formulae in the cluster. From rules generated from all formulae, the rule which gets the maximum values of support and confidence is *Derris scandens* → (77.6%, 77.6%). This herb is usually used as a main component for muscle pain and fatigue. For the clusters, other herbs may be added to produce the secondary indications.

VI. CONCLUSION

This paper showed data mining techniques, i.e., clustering and association rule learning, were useful for finding patterns of herbal formulae with main category of muscle pain and fatigue and several subcategories. With clustering technique, it facilitated herbal experts to find the set of clusters with appropriated subcategories. The association rule learning was applied on the all formulae and formulae on each cluster of the selected set to find the important patterns or rules of combinations. For the future works, herbs which their indications are similar, should be treated as one herb. Other data mining techniques such as classification, should be investigated.

ACKNOWLEDGMENT

This work has been supported by National Science and Technology Development Agency (NSTDA) under project number P-09-00159 as well as the National Electronics and Computer Technology Center (NECTEC) via research grant NT-B-22-MA-17-50-14.

REFERENCES

[1] H. D. Lovell-Smith, "In defence of ayurvedic medicine," *The New Zealand Medical Journal*, vol. 119, no.1234, pp. 1-3, 2006.
[2] Bureau of drug control, available at http://www2.fda.moph.go.th/consumer/drug/dcenter.asp
[3] S. Chomya and V. Ausawakitwiree, "Herbals Composition and use indications Thai traditional medicine for muscle pain," *Journal of Thai Traditional & Alternative Medicine*, vol. 6, no. 2, pp. 50, 2008.
[4] M. Benkhalifa, A. Mouradi, and H. Bouyakhf, "Integrating WordNet knowledge to supplement training data in semi-supervised agglomerative hierarchical clustering for text categorization," *International Journal of Intelligent Systems*, vol. 16, no. 8, pp. 929–947, 2001.
[5] J. Kazama and J. Tsujii, "Maximum entropy models with inequality constraints: A case study on text categorization," *Machine Learning*, vol. 60(1-3), pp. 159–194, 2005.
[6] R. Agrawal, T. Imielinski, and A. N. Swami, "Mining association rules between sets of items in large databases," in P. Buneman and S. Jajodia, editors, *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, pp. 207–216, Washington, D.C., 1993.
[7] A. Inokuchi, T. Washio, and H. Motoda, "An apriori-based algorithm for mining frequent substructures from graph data," In *Proceedings of the PKDD-00, the 4th European Conference on Principles of Data Mining and Knowledge Discovery*, pp. 13–23, Lyon, FR, 2000.

DOI: 01.IJNS.03.02.96